PulseMind: A Multi-Modal Medical Model for Real-World Clinical Diagnosis

Jiao Xu[1,2*], Junwei Liu[2,3*], Jiangwei Lao[2], Qi Zhu[2], Yunpeng Zhao[4], Congyun Jin[2], Shinan Liu[4], Zhihong Lu[2], Lihe Zhang[1†], Xin Chen[5†], Jian Wang[2†], Ping Wang[3†]

[1] Dalian University of Technology
[2] Ant Group
[3] Peking University
[4] University of Hong Kong
[5] City University of Hong Kong
xjmmcome@mail.dlut.edu.cn, wenshuo.ljw@antgroup.com, zhanglihe@dlut.edu.cn

Abstract

Recent advances in medical multi-modal models focus on specialized image analysis like dermatology, pathology, or radiology. However, they do not fully capture the complexity of real-world clinical diagnostics, which involve heterogeneous inputs and require ongoing contextual understanding during patient-physician interactions. To bridge this gap, we introduce PulseMind, a new family of multi-modal diagnostic models that integrates a systematically curated dataset, a comprehensive evaluation benchmark, and a tailored training framework. Specifically, we first construct a diagnostic dataset, MediScope, which comprises 98,000 real-world multi-turn consultations and 601,500 medical images, spanning over 10 major clinical departments and more than 200 sub-specialties. Then, to better reflect the requirements of real-world clinical diagnosis, we develop the PulseMind Benchmark, a multi-turn diagnostic consultation benchmark with a four-dimensional evaluation protocol comprising proactiveness, accuracy, usefulness, and language quality. Finally, we design a training framework tailored for multi-modal clinical diagnostics, centered around a core component named Comparison-based Reinforcement Policy Optimization (CRPO). Compared to absolute score rewards, CRPO uses relative preference signals from multi-dimensional comparisons to provide stable and human-aligned training guidance. Extensive experiments demonstrate that PulseMind achieves competitive performance on both the diagnostic consultation benchmark and public medical benchmarks.

Code — https://github.com/AQ-MedAI/PulseMind

Introduction

In recent years, Visual Language Models (VLMs) (Achiam et al. 2023; Liu et al. 2023; Bai et al. 2025; Zhu et al. 2025; Team et al. 2023) have made remarkable progress across various fields due to their strong capabilities in visual understanding and multi-modal reasoning. These advances have inspired the development of medical multi-modal large models (Li et al. 2023a; Saab et al. 2024; Ayaz et al. 2024; Sellergren et al. 2025; Thawkar et al. 2023; Moor et al. 2023). However, despite this progress, several critical challenges remain insufficiently addressed, especially in real-world clinical applications.

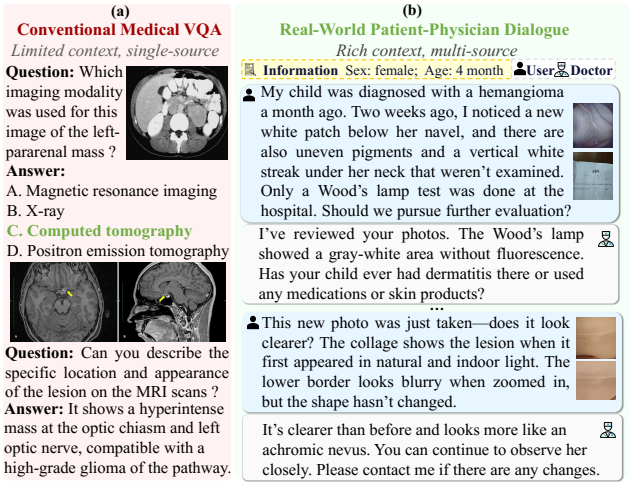


Figure 1: Comparison between (a) conventional multi-modal medical VQA tasks and (b) real-world diagnostic dialogue scenarios.

Compared to conventional medical image analysis (Fig.1(a)), real-world diagnostic consultations (Fig.1(b)) present fundamentally different and significantly complex characteristics. These real-world scenarios involve not only the integration of heterogeneous information from multiple imaging modalities, but also the handling of multi-turn interactions between physicians and patients. This complexity exposes two fundamental gaps in current research:

i) *Limitations in training datasets.* Most existing datasets either focus on visual question answering (VQA) (He et al. 2020; Lau et al. 2018; Chen et al. 2024a) or contain only a single image modality, lacking both the diversity of visual inputs and the multi-turn dialogue context that are critical in

*Equal contribution. Work performed when Jiao Xu was an intern of Ant Group.
†Corresponding authors.

clinical consultations.

ii) *Inadequate evaluation benchmarks.* As a recent study (Xu et al. 2025) pointed out, current medical multi-modal benchmarks fall short of capturing real-world clinical complexity, limiting their ability to evaluate model utility in practical downstream tasks, such as clinical diagnostics. Although a few benchmarks include dialogue cases (Arora et al. 2025), these are mostly limited to pure text and lack integration with medical imagery.

To address these challenges, we propose PulseMind, a new multi-modal diagnostic model that encompasses a systematically curated dataset, an evaluation benchmark, and a tailored training framework.

For the dataset construction, we systematically curated a large-scale multi-modal diagnostic dialogue dataset through the following four key steps: collection, anonymization, expansion, and proofreading. The final dataset consists of 98,000 real-world multi-turn consultation dialogues and 601,500 medical images, spanning over 10 major clinical departments and 200 sub-specialties. It also incorporates a broad spectrum of clinical data types, including laboratory test results, examination reports, prescriptions, medical images, surgical records, and other relevant reports, demonstrating strong clinical diversity and representativeness. For the diagnostic dialogue evaluation, we introduce the PulseMind Benchmark, which features multi-turn scenarios incorporating both plain-text and multi-modal inputs, enabling a comprehensive assessment of diagnostic dialogue capabilities. To better simulate real-world clinical settings, we further develop a GPT-based automatic evaluation framework that assesses model performance across four key dimensions: accuracy, proactiveness, usefulness, and language quality, providing a well-rounded metric for evaluating the quality of diagnostic interactions.

During the model training phase, we first perform supervised fine-tuning (SFT) on medical text and multi-modal data to build domain knowledge and enhance multi-modal understanding. Following this, to better optimize diagnostic responses, we employ a reinforcement learning stage using our proposed Comparison-based Reinforcement Policy Optimization (CRPO) method. The motivation behind CRPO stems from our observation that using absolute numerical scores as rewards often leads to instability and does not align well with human preferences. Humans find it easier to judge which of two responses is better rather than assigning an absolute quality score to a single response. To leverage this insight, CRPO trains the model by comparing its responses pairwise against those from multiple counterpart models. These comparisons evaluate key aspects including proactiveness, accuracy, usefulness, and language quality. By learning from relative preference signals instead of absolute rewards, CRPO provides more stable and human-aligned guidance, ultimately encouraging the model to generate more reliable and higher-quality diagnostic responses.

Experimental results show that PulseMind achieves an average win rate of 76% on the PulseMind Benchmark, indicating its promising diagnostic consultation capabilities. Furthermore, the model achieves competitive performance on 11 public medical question-answering datasets, demonstrating solid generalization capabilities.

In summary, the contributions of this work are three-fold:

- We introduce the *MediScope Dataset*, the first large-scale multi-modal clinical diagnostics dataset featuring rich, real-world multi-turn diagnostic consultations.
- We propose the *PulseMind Benchmark*, the first benchmark for evaluating clinical diagnostic capabilities, providing a comprehensive multi-dimensional assessment of model in multi-turn diagnostic consultations.
- We present *PulseMind*, a medical multi-modal large model specifically designed for real-world clinical diagnostics, achieving competitive performance on both the PulseMind Benchmark and public medical question-answering datasets.

Related Work

General Visual Language Models

Visual Language Models (VLMs) have emerged as a new paradigm for general-purpose artificial intelligence, demonstrating impressive capabilities in multi-modal understanding by leveraging large-scale datasets. The InternVL series (Chen et al. 2024c; Zhu et al. 2025) advances multi-modal learning through the expansion of visual encoders, thereby improving the representation of fine-grained visual details. The Qwen-VL models (Wang et al. 2024a; Bai et al. 2025) build upon strong large language models, showing notable strengths in contextual understanding and instruction following. The OpenAI "o" series (Jaech et al. 2024) and DeepSeek-R1 (Liu et al. 2024; Guo et al. 2025) further push performance boundaries via post-training alignment techniques. Meanwhile, closed-source flagship models such as GPT-4o (Hurst et al. 2024), Gemini 2.5 Pro (Team et al. 2023), and Claude 3.5 (Hu et al. 2024) exhibit strong performance in cross-modal reasoning and dialogue generation, and are often used as reference points for evaluating the capabilities of current VLM systems.

However, despite their success in open-domain tasks, existing models are not readily transferable to the medical domain due to the specialized nature of clinical information (Wu et al. 2023; Nori et al. 2023). In the absence of sufficient medical knowledge, these models often misinterpret critical diagnostic features, resulting in potentially inaccurate or unreliable predictions (Huang et al. 2025; Wang et al. 2024b), which may pose risks to patient safety. These challenges underscore the necessity of developing dedicated multi-modal large models tailored to medical applications.

Medical Visual Language Models

Recent years have witnessed significant advances in medical visual language models, aiming to adapt general-purpose architectures for clinical applications. Early efforts such as LLaVA-Med (Li et al. 2023a) and Med-VLM (Ayaz et al. 2024) enhanced general models by incorporating medical knowledge via large-scale fine-tuning. Subsequently, domain-specific models have been developed for specialized fields including radiology (Thawkar et al. 2023; Zhang et al. 2023a), pathology (Chen et al. 2025b; Zhang et al. 2025),

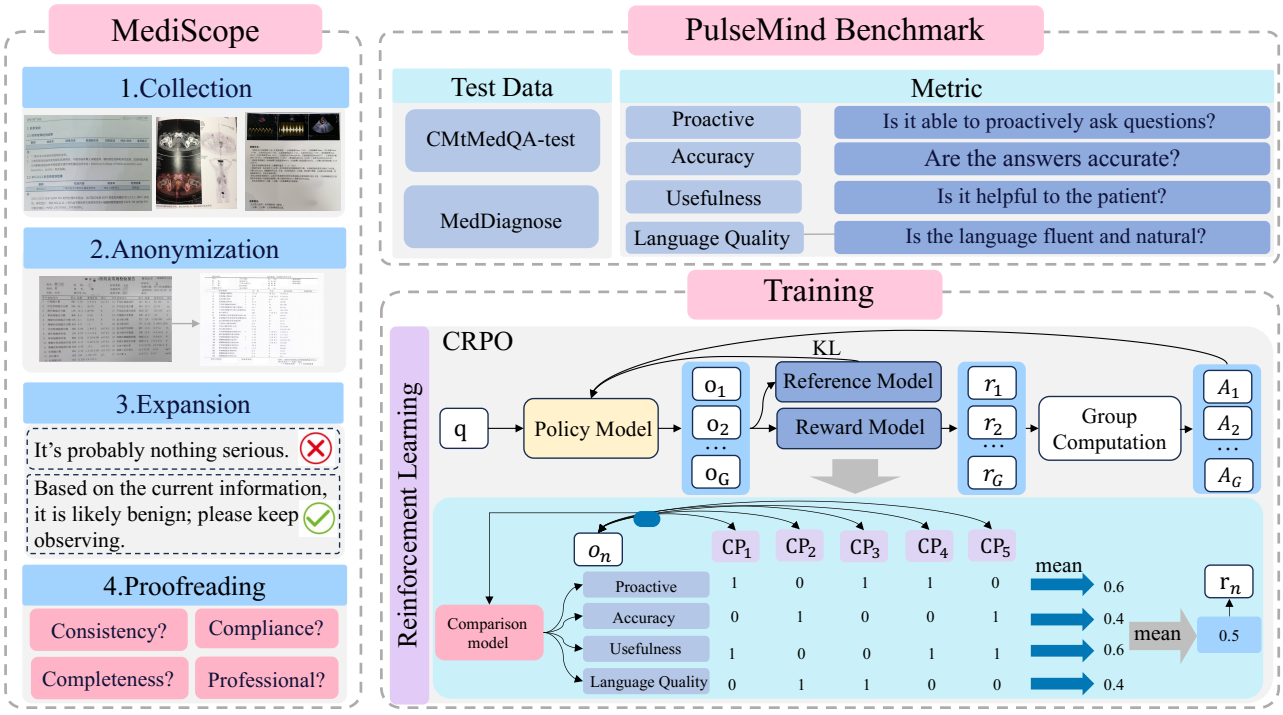


Figure 2: Overview of PulseMind, including dataset construction (MediScope), the PulseMind Benchmark, and CRPO training."CP" denotes the counterpart model.

surgery (Zeng et al. 2025), ophthalmology (Shi et al. 2024), and dermatology (Yan et al. 2024), achieving notable accuracy in their respective domains. More recently, general-purpose medical VLMs such as HuatuoGPT-Vision (Chen et al. 2024a), Med-GEMMA (Sellergren et al. 2025), Lingshu (Xu et al. 2025), and Med-Gemini (Saab et al. 2024) have been developed, trained on large-scale multi-specialty datasets to strike a balance between broad generalization and specialized domain expertise.

Despite these advances, most existing models primarily focus on specialized image analysis and fall short of addressing the complexity inherent in real-world clinical diagnostics. Such diagnostics require the integration of heterogeneous multi-modal inputs and the maintenance of contextual coherence over multi-turn patient-physician interactions. To tackle these challenges, we propose PulseMind, a new multi-modal diagnostic model specifically designed for realistic clinical dialogue scenarios.

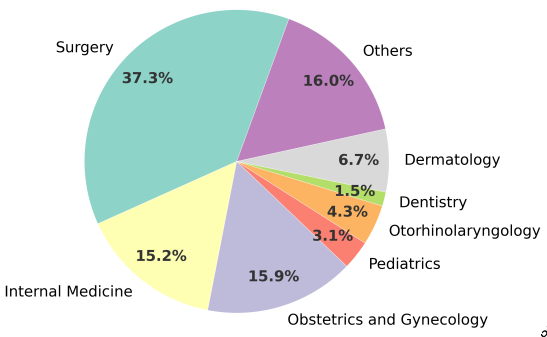
PulseMind

We propose PulseMind, a unified framework tailored for multi-modal physician-patient consultation scenarios. As shown in Fig. 2, our framework consists of three core components. First, we construct a large-scale multi-modal medical dialogue dataset to provide a solid data foundation for model training. Second, we design a comprehensive evaluation benchmark that realistically reflects the key challenges in clinical consultations. Finally, we introduce a Comparison-based Reinforcement Policy Optimization (CRPO) method to conduct reinforcement learning training.

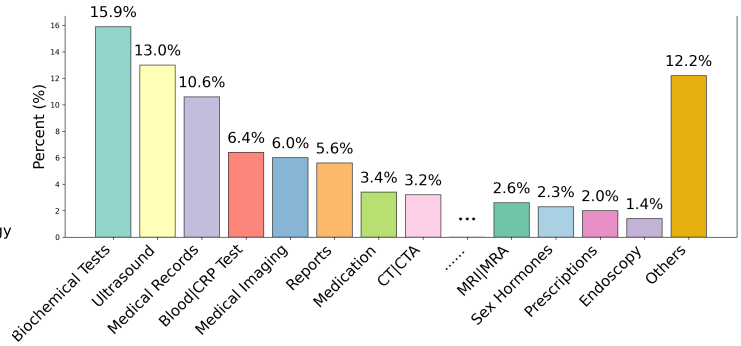
Dataset

Our training data consists of our self-constructed dataset, MediScope, along with a collection of publicly available textual and multi-modal datasets, totaling approximately 792,000 samples.

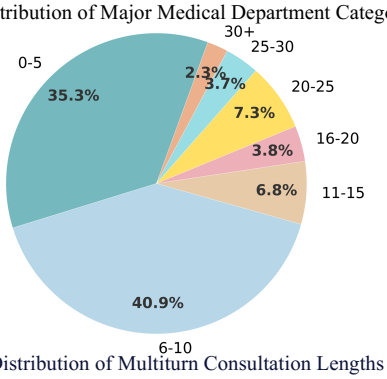
MediScope. We construct a large-scale, heterogeneous multi-modal dataset that captures the complexity of clinical dialogues, referred to as MediScope. The construction process follows a rigorous four-stage pipeline: i) Collection: We collect de-identified data from real-world clinical scenarios, encompassing diverse modalities such as examination reports and multi-turn physician-patient dialogues that authentically capture the complexity and variability of diagnostic processes. ii) Anonymization: We apply advanced Optical Character Recognition (OCR) and Named Entity Recognition (NER) techniques to perform a secondary anonymization check on both text and images, ensuring complete removal of personally identifiable information and compliance with privacy regulations. iii) Expansion: We use large language models (Hurst et al. 2024; Team et al. 2023) to refine and expand physician responses by filtering out meaningless fillers and augmenting clinically relevant content, enhancing the dialogue's completeness, clarity, and coherence without compromising medical accuracy or intent. iv) Proofreading: Medical experts and licensed physicians thoroughly review and refine the expanded dialogues to ensure clinical validity, ethical compliance, and appropriate expression of empathy.



(a) Distribution of Major Medical Department Categories



(b) Distribution of All Medical Image Types



(c) Distribution of Multiturn Consultation Lengths

Patient: <image1,2,3,4>
Doctor: Hello, since both TCT and HPV are abnormal, I suggest you go to a regular hospital for colposcopy to rule out cervical lesions.
Patient: Hello doctor, I have uploaded the colposcopy images. <image5>
Doctor: This colposcopy report lacks any description and is not standardized. I recommend that you get checked at a well-equipped hospital or one with experienced doctors.
Patient: <image6>, this is from May.
Doctor: Was a biopsy done at that time?
Patient: <image7>
Doctor: This is a curettage report, right? This report is normal.
Patient: This sample was taken when my diverticulum was repaired by the doctor. Does this count as a biopsy?
Doctor: This is an endometrial biopsy, not a cervical biopsy.

(d) Case Example of a Multiturn Doctor-Patient Dialogue

Figure 3: Characteristics of the collected multi-turn dialogue training set from four perspectives: (a) Distribution of major departments; (b) Heterogeneous Medical Image Modalities; (c) Distribution of multi-turn dialogue lengths; (d) Example of a multi-turn physician–patient consultation dialogue.

As a result, the dataset comprises 98,000 real-world multi-turn consultations and 601,500 medical images, spanning over 10 major clinical departments and more than 200 sub-specialties. Moreover, it includes a wide variety of data types such as laboratory test results, examination reports, prescriptions, medical images, and surgical records, reflecting strong clinical diversity and representativeness.

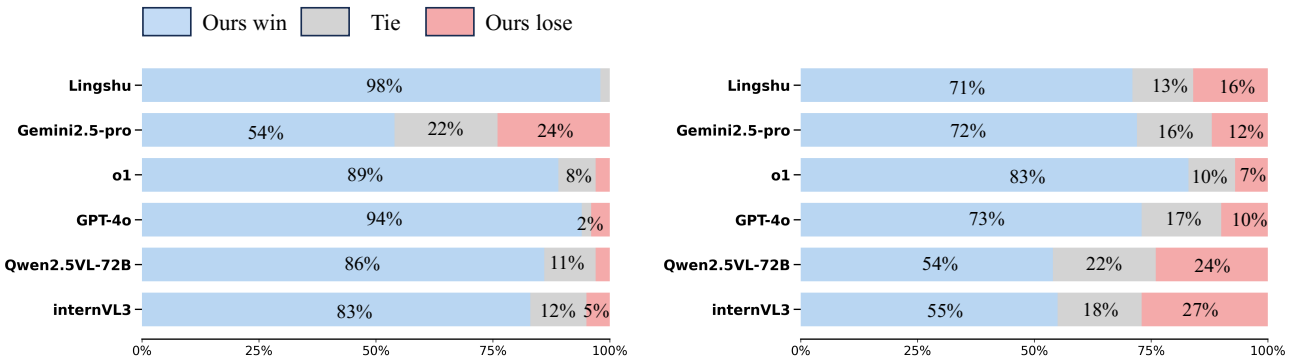
Fig. 3 provides an overview of the dataset's key characteristics across multiple dimensions. Specifically, Fig. 3(a) shows the distribution across a wide range of clinical departments, highlighting the dataset's broad coverage. Fig. 3(b) illustrates the diversity of medical image types, including Ultrasound, Medical Records, CT/MRI, Pathology, and Endoscopy, which supports comprehensive vision-language learning across multiple modalities. Fig. 3(c) shows that the dataset contains a large proportion of multi-turn dialogues, with 40.9% of dialogues containing 6–10 turns and 6% exceeding 20 turns, facilitating long-range dependency modeling and contextual reasoning. Finally, Fig. 3(d) presents a real-world physician-patient conversation that exemplifies the nature of clinical interactions.

Public Datasets. Complementing MediScope, we incorporate a diverse collection of publicly available medical datasets to strengthen the model's capability in traditional medical image analysis. Specifically, we leverage both text-only datasets (Li et al. 2023b; Chen et al. 2024b; Jin et al. 2021; Pal, Umapathi, and Sankarasubbu 2022; Zuo et al. 2025; Hendrycks et al. 2020; Yang et al. 2024a) and multi-modal resources (Yim et al. 2024; He et al. 2020; Zhang et al. 2023b; Liu et al. 2021; Lau et al. 2018; Yue et al. 2024; Zuo et al. 2025).

Evaluation Benchmark

Recent studies (Xu et al. 2025) have pointed out that current medical multi-modal benchmarks do not adequately capture the complexity of real-world clinical scenarios, which limits their ability to evaluate model performance in practical diagnostic tasks. To address this, we introduce the PulseMind Benchmark, which integrates multi-source images and multi-turn dialogue data to realistically simulate clinical workflows and assess diagnostic dialogue capabilities. Furthermore, to situate our results within the broader field and enable fair comparison with existing methods, we also evaluate our models on standard medical QA benchmarks.

PulseMind Benchmark To comprehensively evaluate model capabilities in real-world medical consultation scenarios, we propose the PulseMind Benchmark, which integrates three core components: (i) a multi-source dataset covering both text-based and multi-modal multi-turn diagnostic dialogues; (ii) clinically driven multi-dimensional evaluation metrics; and (iii) a human-aligned relative scoring strategy. Together, these components establish a unified framework to systematically and rigorously assess model performance in authentic clinical environments.



(a) Evaluation results on the MedDiagnose dataset

(b) Evaluation results on the CMtMedQA-test dataset

Figure 4: Win rates of our model against six baseline methods on the PulseMind Benchmark

Method	Muilti-modal QA							Text-only QA			
	MMMU	VQA-RAD	PMC-VQA	SLAKE	PathVQA	DermaVQA	MedXQA-MM	MMLU	MedMCQA	MedQA	MedXQA-text
Proprietary Models											
GPT-4o	57.3	71.2	55.2	67.4	55.5	35.0	22.3	88.7	73.5	55.7	22.5
o1	57.8	63.0	54.5	69.9	57.3	43.0	49.7	91.6	82.7	86.6	48.9
Gemini2.5-pro	49.3	70.5	55.5	75.8	55.4	39.0	39.5	89.8	68.6	85.6	24.3
Open-source Models (∼72B)											
InternVL3-78B	69.1	73.6	56.6	77.4	51.0	37.0	27.4	83.0	66.1	93.3	18.5
Qwen2.5VL-72B	66.4	80.3	59.3	78.3	42.3	34.0	27.6	88.3	67.2	91.3	16.1
PulseMind-72B	**69.4**	**87.1**	**70.3**	**85.6**	**64.9**	**42.0**	**36.7**	**88.7**	**71.3**	**94.8**	**29.8**
Open-source Models (∼32B)											
InternVL3-38B	**65.2**	65.4	56.6	72.7	51.0	31.0	25.2	82.8	64.9	73.5	16.0
Qwen2.5VL-32B	62.8	73.8	54.5	71.2	41.9	25.0	25.2	83.2	63.0	71.6	15.6
LLAVA-med-34B	48.9	58.6	44.4	67.3	48.8	13.0	16.4	74.7	52.2	63.5	14.1
HuatuoGPT-vision-34B	54.3	61.4	56.6	69.5	44.4	21.0	17.3	80.8	63.6	57.4	16.0
Lingshu-32B	62.3	76.5	57.9	**89.2**	**65.9**	17.0	**30.9**	84.7	66.1	74.7	**22.7**
PulseMind-32B	64.6	**83.2**	**68.1**	81.5	62.0	**32.0**	29.6	**85.6**	**66.4**	**92.9**	21.5

Table 1: Performance comparison on medical QA benchmarks. The top two results highlighted with **bold** and underlined fonts, respectively. Rows with gray background indicate our PulseMind models. "MMMU" refers to MMMU Health & Medicine, "MedXQA" refers to MedXpertQA, and "MMLU" refers to MMLU clinical topics.

Benchmark Composition. The PulseMind Benchmark combines two datasets, covering both text-only and multi-modal consultation scenarios with over 1,200 samples in total. The multi-modal consultation set, named MedDiagnose, was constructed by us and contains 237 samples collected from patient cases, featuring images alongside expert-verified dialogues. For text-based dialogue understanding, we expand the original CMtMedQA-test dataset to include multi-turn reasoning, resulting in 1,000 samples.

Multi-dimensional Evaluation Protocol. Evaluating physician-patient consultation quality requires more than correctness, involving clinical reasoning, contextual interaction, and communicative clarity (Yang et al. 2024b; Chen et al. 2025a). To capture these aspects, we design an evaluation protocol based on four key dimensions:

- Proactiveness, which assesses whether the model actively inquires about missing but critical information, emulating the diagnostic behavior of experienced physicians.
- Accuracy, which verifies whether the diagnostic suggestions are medically sound and free from factual errors or inappropriate reasoning.
- Usefulness, which measures the practical value of the response, including clarity, actionability, and relevance to the patient's concerns.
- Language Quality, which evaluates fluency, professionalism, and overall communicative effectiveness.

These four dimensions jointly capture both clinical content quality and the effectiveness of medical communication, providing a holistic view of model performance.

Evaluation Strategy. We adopt GPT-4 as an automatic evaluator, following recent studies (Zheng et al. 2023; Sun et al. 2023). For each input prompt, model responses are compared against those of multiple baselines across the four evaluation dimensions mentioned above. The outcomes are categorized as win, tie, or loss, and the win rate is used as the primary evaluation metric.

General Medical QA Benchmarks. In addition to consultation evaluation, we also assess model performance on traditional medical question answering tasks in both multi-modal and text-only settings, using 11 benchmark datasets.

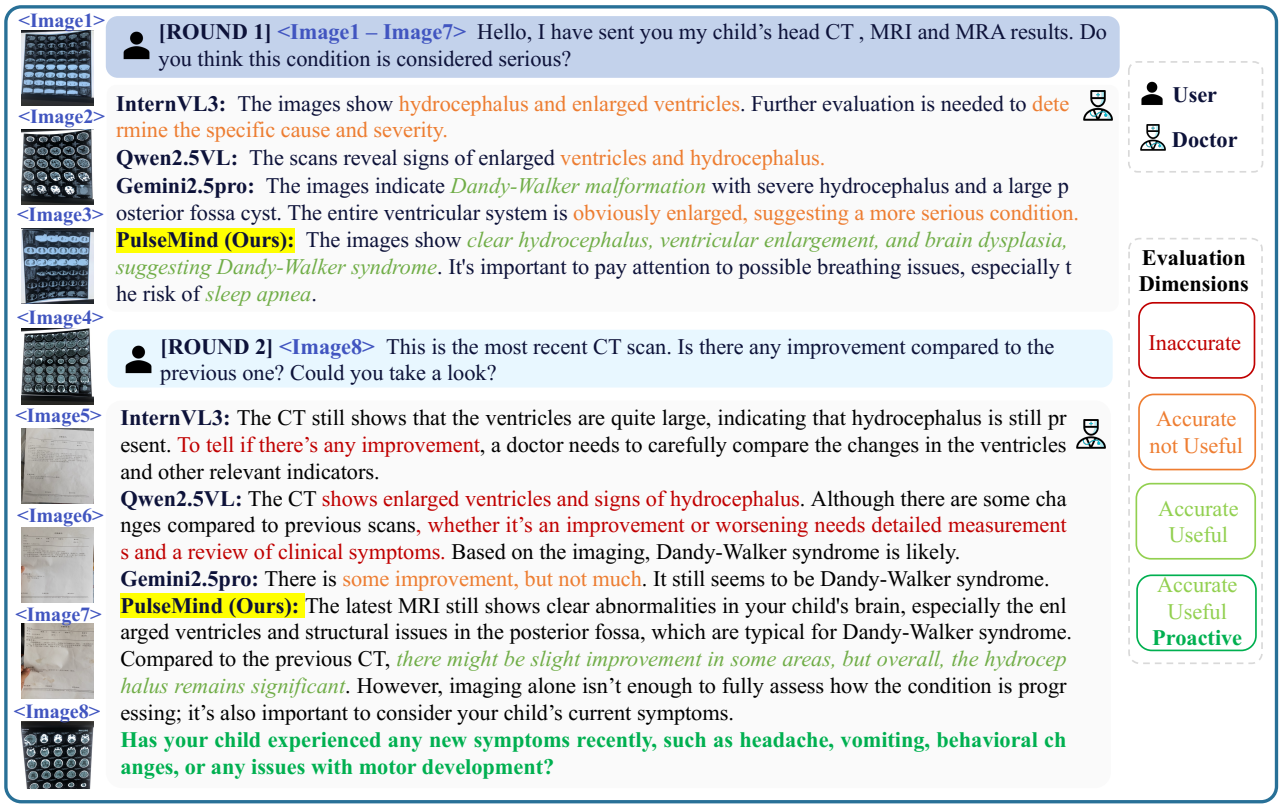


Figure 5: Illustrative cases of six models on the PulseMind Benchmark. Representative response quality is color-coded, as indicated by the tags on the right.

Benchmark Composition. For multi-modal QA, we select datasets that cover general medical knowledge (Yue et al. 2024; Zhang et al. 2023b), clinical reasoning (Zuo et al. 2025; Liu et al. 2021), and specialized domains (He et al. 2020; Lau et al. 2018; Yim et al. 2024). For text-based QA, we use large-scale medical examinations (Jin et al. 2021; Pal, Umapathi, and Sankarasubbu 2022), MMLU clinical topics (Hendrycks et al. 2020), and the text-only subset of MedXpertQA (Zuo et al. 2025) to evaluate the model's advanced clinical reasoning abilities.

Evaluation Strategy. For general medical QA benchmarks, we follow the official evaluation protocols to ensure consistency and comparability with existing methods. For multiple-choice questions, a two-stage evaluation is adopted: exact rule-based matching first, and if no match is found, the answer with the highest semantic similarity is selected. For open-ended questions, GPT-4 is used for automated evaluation to judge the semantic consistency between the model's response and the reference answer.

Training Framework

The training includes supervised fine-tuning on medical text and multi-modal data, followed by reinforcement learning tailored to diagnostic dialogue.

Supervised Fine-Tuning. First, we train the model on Huatuo26M (Li et al. 2023b), to inject domain-specific knowledge and enhance its language understanding and clinical reasoning capabilities. Second, we further train the model on MediScope and public datasets to unlock its multimodal capabilities and multi-turn medical dialogue capabilities, thereby enhancing its ability to process clinical dialogues.

Reinforcement Learning with CRPO. After SFT, the model acquires the ability to perform routine diagnostic consultations. To further enhance its human-aligned performance in real-world clinical consultation scenarios, we employ a RL approach to optimize the model across four key dimensions: proactiveness, accuracy, usefulness, and language quality. Popular RL methods like GRPO(Shao et al. 2024), which rely on absolute scores as rewards, face inherent limitations in clinical dialogue scenarios. First, model-based absolute scores tend to be unstable and subjective. Second, whether model-based or rule-based, absolute scores often obscure differences among top-performing models, limiting their discriminative power. To address these issues, we replace absolute score rewards with comparison-based relative rewards, and propose a Comparison-based Reinforcement Policy Optimization algorithm (CRPO) that facilitates a stable and human-aligned optimization process.

Specifically, as illustrated in the lower-right part of Fig. 2, given a sampled query q, the policy model generates a set of candidate responses $\{o_1, o_2, \ldots, o_G\}$. These responses are then evaluated by a reward model, which con-

Model	Proact.	Acc.	Use.	Lang.	Avg.
Lingshu	3.60	4.15	4.05	4.58	4.10
Gemini2.5-pro	3.91	4.35	4.30	4.69	<u>4.31</u>
o1	3.70	4.43	4.14	4.60	4.22
GPT-4o	3.48	4.06	3.97	4.52	4.01
Qwen2.5VL-72B	3.81	4.30	4.22	4.59	4.23
InternVL3-78B	3.74	4.36	4.16	4.71	4.24
PulseMind	3.92	4.47	4.26	4.763	**4.35**

Table 2: Absolute scores assigned to seven models across four evaluation dimensions. The top two results are highlight with **bold** and <u>underlined</u> fonts, respectively.

sists of a comparison model and five counterpart models $\{CP_1, \ldots, CP_5\}$. The comparison model assesses each candidate response against those generated by the counterpart models across four key evaluation dimensions: proactiveness, accuracy, usefulness, and language quality.

For each candidate o_g, the comparison model determines whether it performs better than each counterpart CP_c on each evaluation dimension d. Concretely, the comparison model assigns a binary score:

$$r_{g,c,d} = \begin{cases} 1, & \text{if } o_g \succ CP_c \text{ on dimension } d, \\ 0, & \text{otherwise,} \end{cases}$$

where $r_{g,c,d}$ indicates whether the g-th candidate response performs better than the c-th counterpart model on the d-th dimension. The reward for the candidate o_g is obtained by averaging over all counterpart models and evaluation dimensions:

$$R_g = \frac{1}{C \times D} \sum_{c=1}^{C} \sum_{d=1}^{D} r_{g,c,d},$$

where $C = 5$ is the number of counterpart models and $D = 4$ is the number of evaluation dimensions. Following this, the advantage and final loss computation follow the same procedure as in GRPO.

Discussion. To gain a clearer understanding of the differences between relative and absolute evaluation strategies, we conducted experiments examining their scoring behaviors and evaluation reliability. First, we analyze how absolute scoring performs in distinguishing different models. Second, we assess the reliability of the two strategies by comparing their consistency with human expert evaluations.

To evaluate the distribution and discriminative capacity of the absolute scoring strategy, we rated seven models using a 5-point scale across four evaluation dimensions: Proactiveness (Proact.), Accuracy (Acc.), Usefulness (Use.), and Language Quality (Lang.). As shown in Tab. 2, all models received relatively high and closely clustered scores, with average ratings ranging from 4.01 to 4.35. Such small score differences make it difficult to effectively differentiating the model performance.

To validate the effectiveness of the relative scoring approach, we randomly sampled 10% of the evaluation outputs from both the absolute and relative scoring strategies and asked 50 medical experts to verify whether they agreed with each judgment. Based on their agreements, we computed the consistency rates between GPT-based evaluations and expert assessments across four dimensions as well as overall. As shown in Tab. 3, the relative scoring strategy achieves an average consistency of 86.1% on the PulseMind Benchmark, whereas the absolute scoring strategy reaches only 51.5%, demonstrating the superior reliability of relative evaluation.

		Proact.	Acc.	Use.	Lang.	All
Abs	MedDi	53.7%	48.6%	48%	64.5%	42.1%
	CMt	66.7%	58.1%	61.5%	73.4%	60.9%
Rel	MedDi	96%	88%	86.8%	92.2%	84.2%
	CMt	98.4%	81.1%	94.1%	98.2%	87.9%

Table 3: Consistency between GPT-based evaluations and human expert judgments under absolute (Abs) and relative (Rel) scoring strategies. MedDi and CMt denote the two subsets of the PulseMind Benchmark.

Experiments

Implementation Details

We build on Qwen2.5-VL-72B and Qwen2.5-VL-32B. To achieve efficient parameter fine-tuning, we adopt a low-rank adaptation (LoRA) strategy by injecting a rank-64 adaptation matrix into the Transformer layer to freeze the base model while only training the newly added parameters. Our training process is implemented on a cluster of 128 NVIDIA A100 GPUs and integrates a set of advanced optimization technology stacks. The technology stack is based on HuggingFace Transformers and PEFT libraries, combined with DeepSpeed ZeRO-3 to manage GPU memory, and BF16 mixed precision is enabled to accelerate calculations. The model is optimized using the AdamW optimizer, the learning rate is controlled by the cosine annealing scheduler, and a dropout rate of 0.1 is set to enhance generalization ability.

To evaluate performance, we benchmark against a wide range of models. General-purpose proprietary MLLMs include GPT-4o (Achiam et al. 2023), o1 (Jaech et al. 2024), and Gemini 2.5 Pro (Team et al. 2023). Medical-specialized models include LLaVA-Med (Li et al. 2023a), HuatuoGPT-Vision (Chen et al. 2024a), and Lingshu (Xu et al. 2025). For broader comparison, we also include open-source general-purpose MLLMs such as InternVL3 (Zhu et al. 2025) and Qwen2.5-VL (Bai et al. 2025).

State-of-the-Art Comparisons

The evaluation encompasses the PulseMind Benchmark as well as existing medical QA tasks, offering a comprehensive assessment of the models.

PulseMind Benchmarks. As shown in Fig. 4, we evaluate the PulseMind model against leading models on the two subsets of the PulseMind Benchmark: MedDiagnose and cMtMedQA-test.

As shown in Fig. 4(a), on the multi-modal MedDiagnose benchmark, our PulseMind model demonstrates superior performance. Against proprietary general-purpose models, it achieves win rates of 94% against GPT-4o, 89%

	Public	Public+MediScope
PulseMind-B	26.4	65.2
MMMU	67.3	68.1
VQA-RAD	86.6	86.9
SLAKE	84.7	85.3
MedXQA-MM	34.9	36.5

(a) Dataset

	SFT	SFT+RL
PulseMind-B	65.2	76.0
MMMU	68.1	69.4
VQA-RAD	86.9	87.1
SLAKE	85.3	85.6
MedXQA-MM	36.5	36.6

(b) Training Stages

	GRPO	CRPO
PulseMind-B	54.7	76.0
MMMU	66.7	69.4
VQA-RAD	86.9	87.1
SLAKE	85.2	85.6
MedXQA-MM	36.1	36.6

(c) CRPO *v.s.* GRPO

Table 4: **PulseMind Ablation Experiments**. "PulseMind-B" denotes our PulseMind Evaluation Benchmark. "Public" indicates models trained on public datasets, while "MediScope" refers to our dataset. "SFT" and "RL" denote using supervised fine-tuning or reinforcement learning, respectively. "GRPO" and "CRPO" represent different reinforcement learning strategies.

against o1, and 54% against Gemini 2.5-Pro. For open-source general-purpose models such as Qwen2.5VL-72B and InternVL3, PulseMind attains win rates of 86% and 83%, respectively. When compared to the domain-specific medical model Lingshu, PulseMind shows a clear advantage with a win rate of 98%. These results underscore PulseMind's capabilities in complex clinical scenarios. As shown in Fig. 4(b), on the expanded CMtMedQA-test benchmark, PulseMind maintains superior performance. Against proprietary general-purpose models, it achieves win rates of 83% with o1, 73% with GPT-4o, and 72% with Gemini 2.5-Pro. For open-source general-purpose models, it secures 54% against Qwen2.5VL-72B and 55% against InternVL3. When compared with the domain-specific medical model Lingshu, PulseMind attains a win rate of 71%. These results demonstrate PulseMind's robust generalization capability, even in text-only diagnostic tasks that emphasize broad medical knowledge. In summary, PulseMind exhibits robust performance across both multi-modal and text-only multi-turn clinical consultation tasks, showcasing high adaptability to real-world clinical scenarios. Representative cases are shown in Fig. 5 to illustrate model behaviors.

Medical QA Benchmarks. As shown in Tab. 1, PulseMind demonstrates leading performance on most medical QA benchmarks. Among open-source models with approximately 32B parameters, PulseMind-32B generally performs better than mainstream models such as Lingshu-32B and HuatuoGPT-vision-34B across most benchmarks. In comparisons involving larger-scale models, PulseMind-72B achieves the best results across all 11 benchmarks, surpassing peer open-source models and outperforming closed-source models in multiple tasks.

Ablation Study

We conduct ablation studies using the PulseMind-72B model on the proposed PulseMind Benchmark as well as representative medical QA benchmarks to assess the impact of key design choices. The results are summarized in Tab. 4.

Dataset. As shown in Tab. 4(a), our curated multi-turn, multi-modality dataset, MediScope, brings substantial performance improvements on the PulseMind Benchmark. The average win rate (against six baseline models in Fig. 4) increases significantly from 26.4% to 65.2%, highlighting the importance of constructing a heterogeneous, multi-source dataset grounded in realistic multi-turn diagnostic dialogues. In addition, medical QA tasks also benefit from this dataset. For instance, the accuracy on MedXpertQA increases from 34.9% to 36.5%. Although the improvements on these public benchmarks are relatively modest, they demonstrate the strong generalization capability of MediScope beyond its primary target scenario.

Training Stages. As shown in Tab. 4(b), incorporating reinforcement learning further improves the model's capability, raising the average win rate on PulseMind Benchmark from 65.2% to 76.0%. This highlights the effectiveness of reward-guided policy optimization in complex multi-turn dialogues. On medical QA benchmarks, reinforcement learning yields marginal gains, suggesting that supervised fine-tuning already sufficiently exploits the learning signal for these relatively simple tasks.

CRPO *v.s.* GRPO. Table 4(c) compares our CRPO approach with the conventional GRPO strategy. CRPO demonstrates superior performance to GRPO on the PulseMind Benchmark, increasing the average win rate from 54.7% to 76.0%. It also shows improvements on other medical QA benchmarks. In particular, it achieves a 2.7% improvement on the MMMU Health & Medicine dataset.These results demonstrate the advantage of comparison-based relative rewards over absolute score-based rewards.

Conclusion

This work presents PulseMind, a multi-modal medical model for real-world clinical diagnosis. It includes a large-scale diagnostic dataset (MediScope), a diagnostic evaluation benchmark (PulseMind Benchmark), and a Comparison-based Reinforcement Policy Optimization method (CRPO). PulseMind demonstrates competitive performance on both the proposed benchmark and public medical benchmarks. We hope PulseMind can serve as a solid foundation for practical diagnostic dialogue applications.

Limitation. PulseMind delivers strong multi-modal diagnostic capabilities, yet certain limitations persist. First, its ability to process specialized data formats, such as 3D medical imaging and other high-dimensional clinical modalities, remains limited. Second, training models demands substantial computational resources and considerable time, which may constrain their use in resource-limited environments.

Acknowledgements

This work is supported by Ant Group Research Intern Program and National Natural Science Foundation of China under Grant 62431004.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arora, R. K.; Wei, J.; Hicks, R. S.; Bowman, P.; Quiñonero-Candela, J.; Tsimpourlas, F.; Sharman, M.; Shah, M.; Vallone, A.; Beutel, A.; et al. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Ayaz, M.; Khan, M.; Saqib, M.; Khelifi, A.; Sajjad, M.; and Elsaddik, A. 2024. Medvlm: Medical vision-language model for consumer devices. *IEEE Consumer Electronics Magazine*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, J.; Gui, C.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G. H.; Wang, X.; Zhang, R.; Cai, Z.; Ji, K.; et al. 2024a. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Chen, J.; Gui, C.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G. H.; Wang, X.; Zhang, R.; Cai, Z.; Ji, K.; et al. 2024b. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Chen, X.; Xiang, J.; Lu, S.; Liu, Y.; He, M.; and Shi, D. 2025a. Evaluating large language models and agents in healthcare: key challenges in clinical applications. *Intelligent Medicine*.

Chen, Y.; Wang, G.; Ji, Y.; Li, Y.; Ye, J.; Li, T.; Hu, M.; Yu, R.; Qiao, Y.; and He, J. 2025b. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5134–5143.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

He, X.; Zhang, Y.; Mou, L.; Xing, E.; and Xie, P. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hu, S.; Ouyang, M.; Gao, D.; and Shou, M. Z. 2024. The dawn of gui agent: A preliminary case study with claude 3.5 computer use. *arXiv preprint arXiv:2411.10323*.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 1–55.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 6421.

Lau, J. J.; Gayen, S.; Ben Abacha, A.; and Demner-Fushman, D. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 1–10.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 28541–28564.

Li, J.; Wang, X.; Wu, X.; Zhang, Z.; Xu, X.; Fu, J.; Tiwari, P.; Wan, X.; and Wang, B. 2023b. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; and Wu, X.-M. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging*, 1650–1654.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 34892–34916.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E. P.; and Rajpurkar, P. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health*, 353–367.

Nori, H.; King, N.; McKinney, S. M.; Carignan, D.; and Horvitz, E. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, 248–260.

Saab, K.; Tu, T.; Weng, W.-H.; Tanno, R.; Stutz, D.; Wulczyn, E.; Zhang, F.; Strother, T.; Park, C.; Vedadi, E.; et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Sellergren, A.; Kazemzadeh, S.; Jaroensri, T.; Kiraly, A.; Traverse, M.; Kohlberger, T.; Xu, S.; Jamil, F.; Hughes, C.; Lau, C.; et al. 2025. MedGemma Technical Report. *arXiv preprint arXiv:2507.05201*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shi, D.; Zhang, W.; Yang, J.; Huang, S.; Chen, X.; Yusufu, M.; Jin, K.; Lin, S.; Liu, S.; Zhang, Q.; et al. 2024. EyeCLIP: A visual-language foundation model for multi-modal ophthalmic image analysis. *arXiv preprint arXiv:2409.06644*.

Sun, Z.; Shen, Y.; Zhou, Q.; Zhang, H.; Chen, Z.; Cox, D.; Yang, Y.; and Gan, C. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 2511–2565.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Thawkar, O.; Shaker, A.; Mullappilly, S. S.; Cholakkal, H.; Anwer, R. M.; Khan, S.; Laaksonen, J.; and Khan, F. S. 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, X.; Pan, J.; Ding, L.; and Biemann, C. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*.

Wu, C.; Lei, J.; Zheng, Q.; Zhao, W.; Lin, W.; Zhang, X.; Zhou, X.; Zhao, Z.; Zhang, Y.; Wang, Y.; et al. 2023. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909*.

Xu, W.; Chan, H. P.; Li, L.; Aljunied, M.; Yuan, R.; Wang, J.; Xiao, C.; Chen, G.; Liu, C.; Li, Z.; et al. 2025. Lingshu: A Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning. *arXiv preprint arXiv:2506.07044*.

Yan, S.; Yu, Z.; Primiero, C.; Vico-Alonso, C.; Wang, Z.; Yang, L.; Tschandl, P.; Hu, M.; Tan, G.; Tang, V.; et al. 2024. A general-purpose multimodal foundation model for dermatology. *arXiv preprint arXiv:2410.15038*.

Yang, S.; Zhao, H.; Zhu, S.; Zhou, G.; Xu, H.; Jia, Y.; and Zan, H. 2024a. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, 19368–19376.

Yang, S.; Zhao, H.; Zhu, S.; Zhou, G.; Xu, H.; Jia, Y.; and Zan, H. 2024b. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, 19368–19376.

Yim, W.-w.; Fu, Y.; Sun, Z.; Abacha, A. B.; Yetisgen, M.; and Xia, F. 2024. Dermavqa: A multilingual visual question answering dataset for dermatology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 209–219.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.

Zeng, Z.; Zhuo, Z.; Jia, X.; Zhang, E.; Wu, J.; Zhang, J.; Wang, Y.; Low, C. H.; Jiang, J.; Zheng, Z.; et al. 2025. SurgVLM: A Large Vision-Language Model and Systematic Evaluation Benchmark for Surgical Intelligence. *arXiv preprint arXiv:2506.02555*.

Zhang, W.; Zhang, P.; Guo, J.; Cheng, T.; Chen, J.; Zhang, S.; Zhang, Z.; Yi, Y.; and Bu, H. 2025. Patho-R1: A Multimodal Reinforcement Learning-Based Pathology Expert Reasoner. *arXiv preprint arXiv:2505.11404*.

Zhang, X.; Wu, C.; Zhang, Y.; Xie, W.; and Wang, Y. 2023a. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 4542.

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2023b. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 46595–46623.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Zuo, Y.; Qu, S.; Li, Y.; Chen, Z.; Zhu, X.; Hua, E.; Zhang, K.; Ding, N.; and Zhou, B. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.